Statistical Privacy For Privacy Preserving Information Sharing

Johannes Gehrke

Cornell University

http://www.cs.cornell.edu/johannes

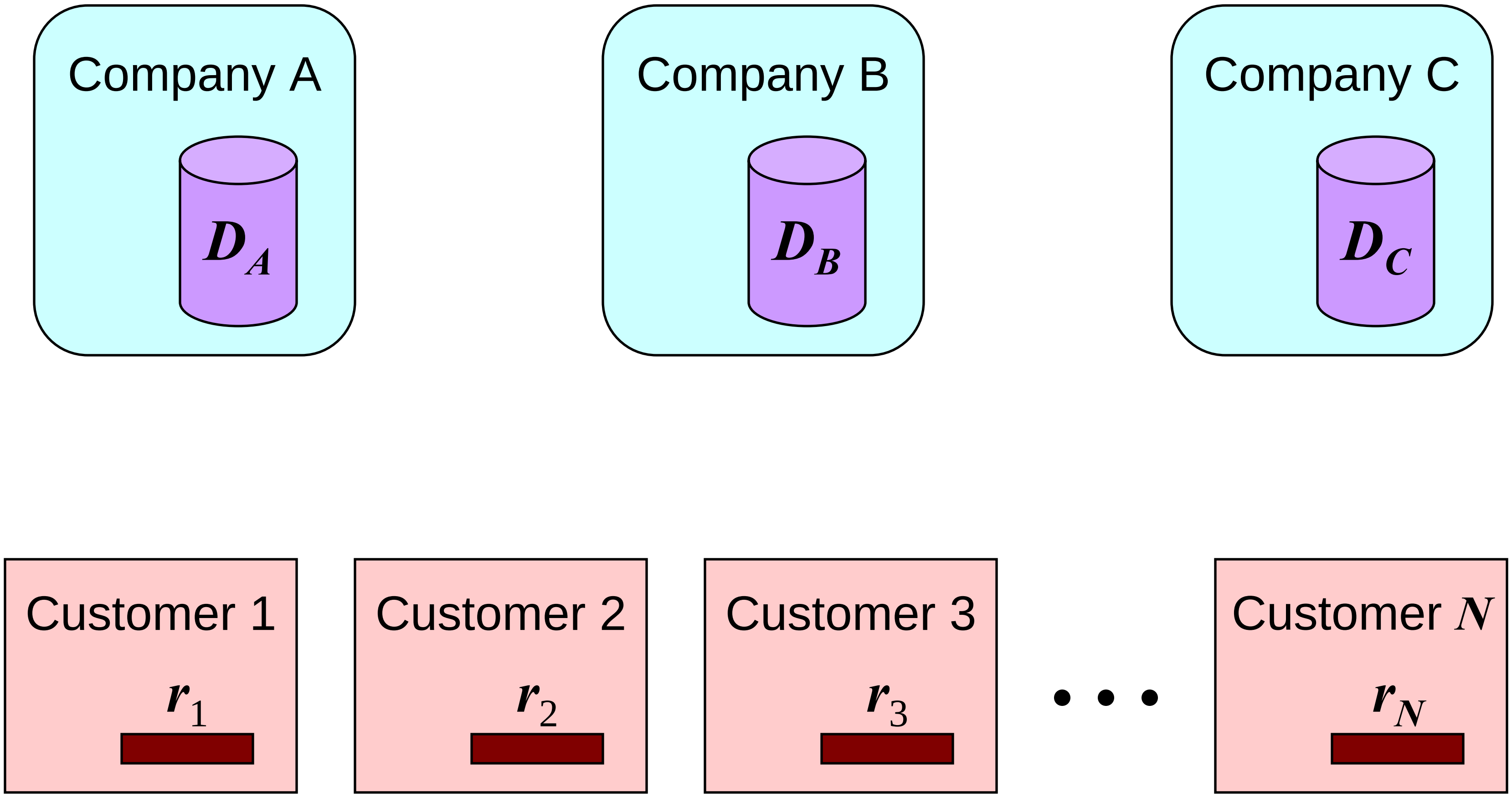
Joint work with: Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal

CORNELL

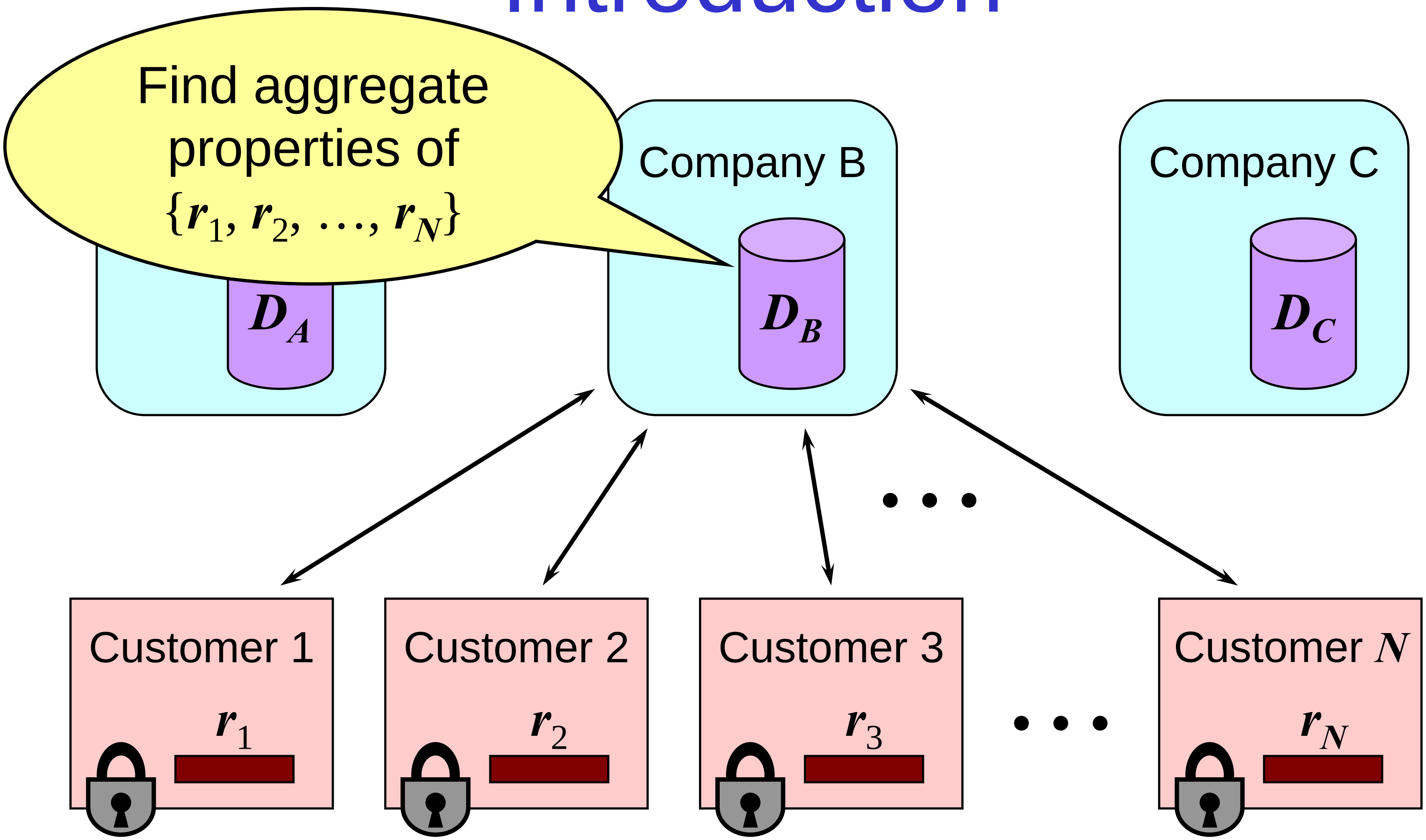
Introduction

- To operate, an e-business needs to query data owned by clients or other businesses
- The owners are concerned about privacy of their data, they will not ship all data to one server
- We want algorithms that efficiently evaluate multi-party queries while disclosing as little extra information as possible.

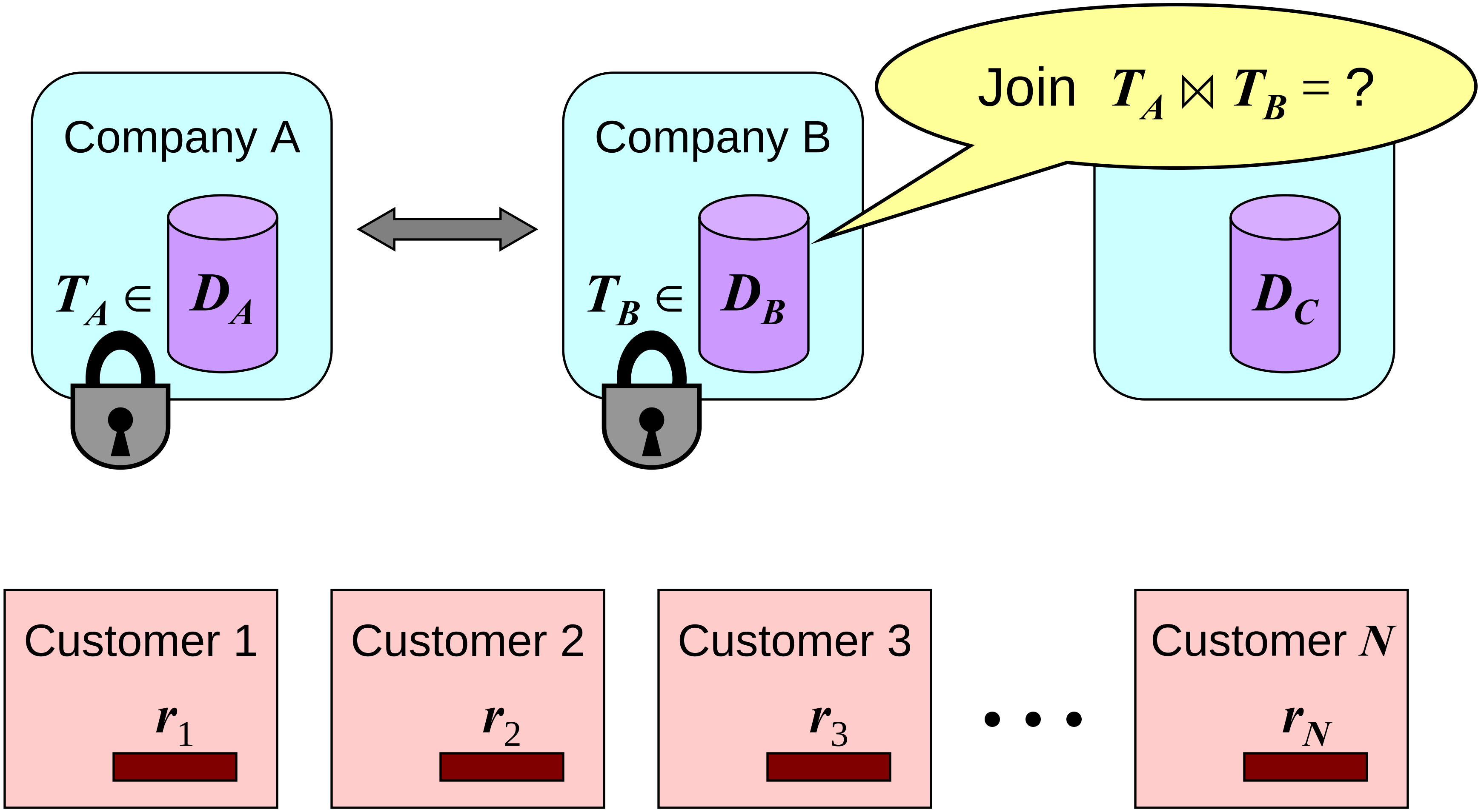
Introduction



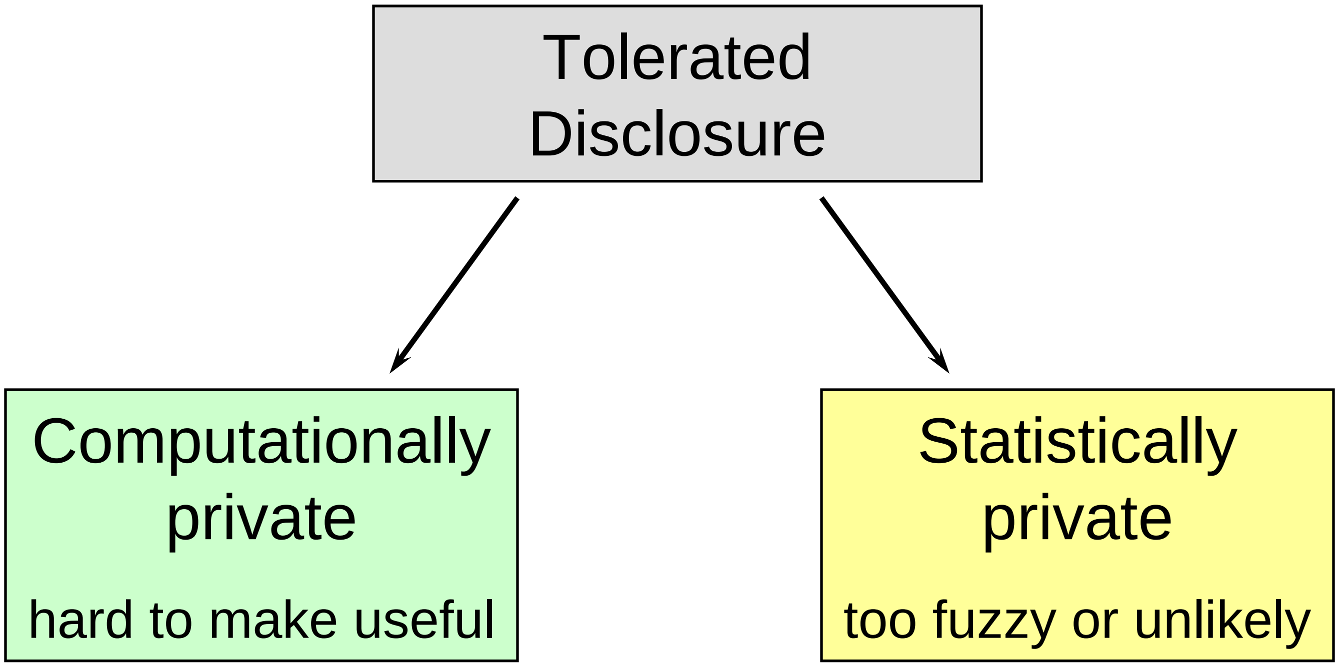
Introduction



Introduction



Privacy and Disclosure



Privacy and Disclosure

Tolerated Disclosure

Most studied

Computationally private

hard to make useful

Statistically private

too fuzzy or unlikely

- Yao [1986]: Any two-party data operation can be made computationally private, if the operation is converted into a Boolean circuit.

Privacy and Disclosure

Tolerated Disclosure

Today's Talk

Computationally private

hard to make useful

Statistically private

too fuzzy or unlikely

Talk Outline

- Introduction
- Randomization and privacy for association rules
- Amplification: upper bound on breaches
- Experimental results
- Conclusion

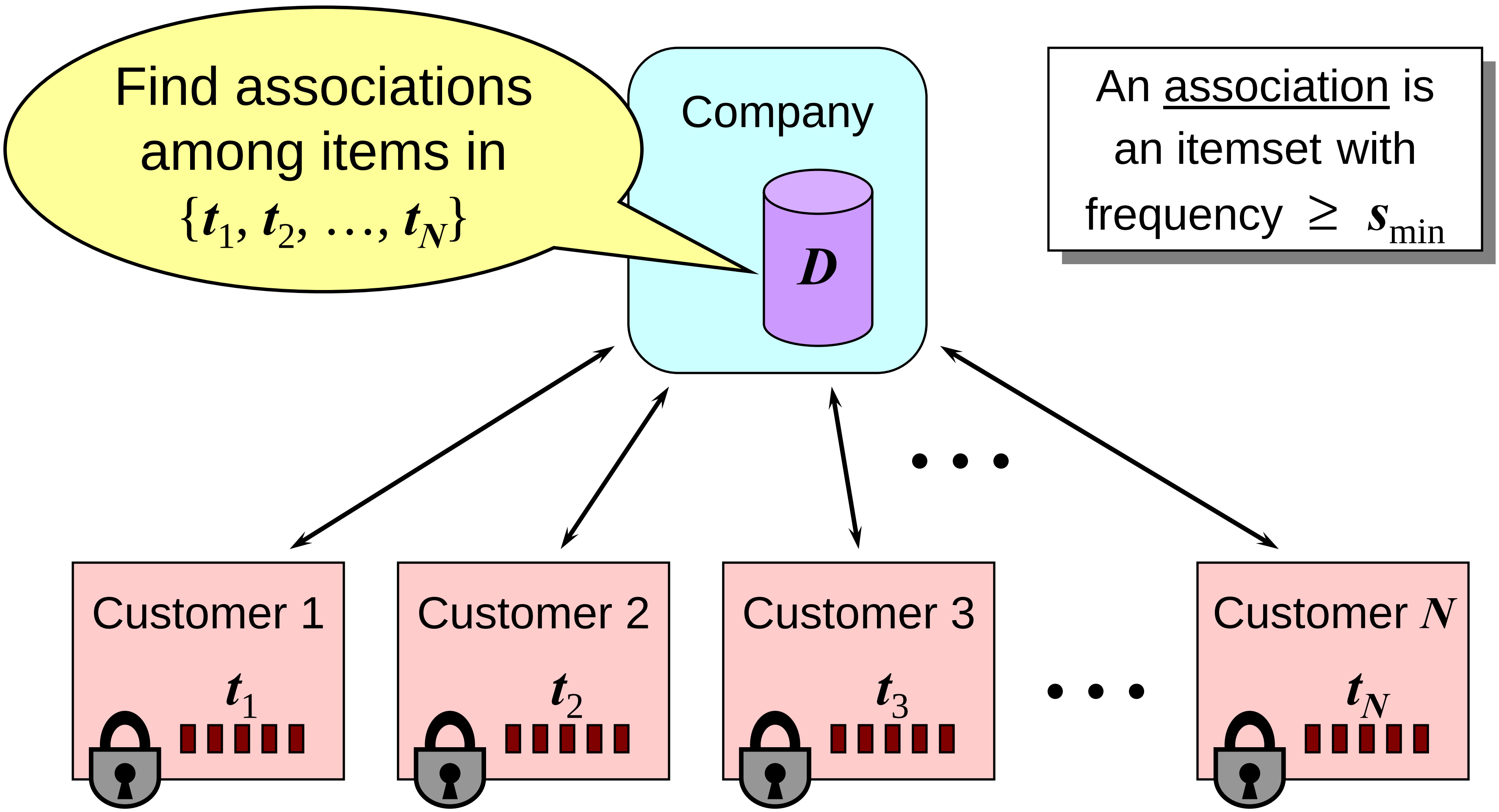
CORNELL

Privacy Preserving Associations

- We have one server and many clients
- Each client has a private transaction (a set of items)
  - Example: product preferences
- The server wants to find frequent subsets of items
  - (aggregate statistical information)
- Each client wants to hide its transaction from the server

CORNELL

Privacy Preserving Associations



Privacy Preserving Associations

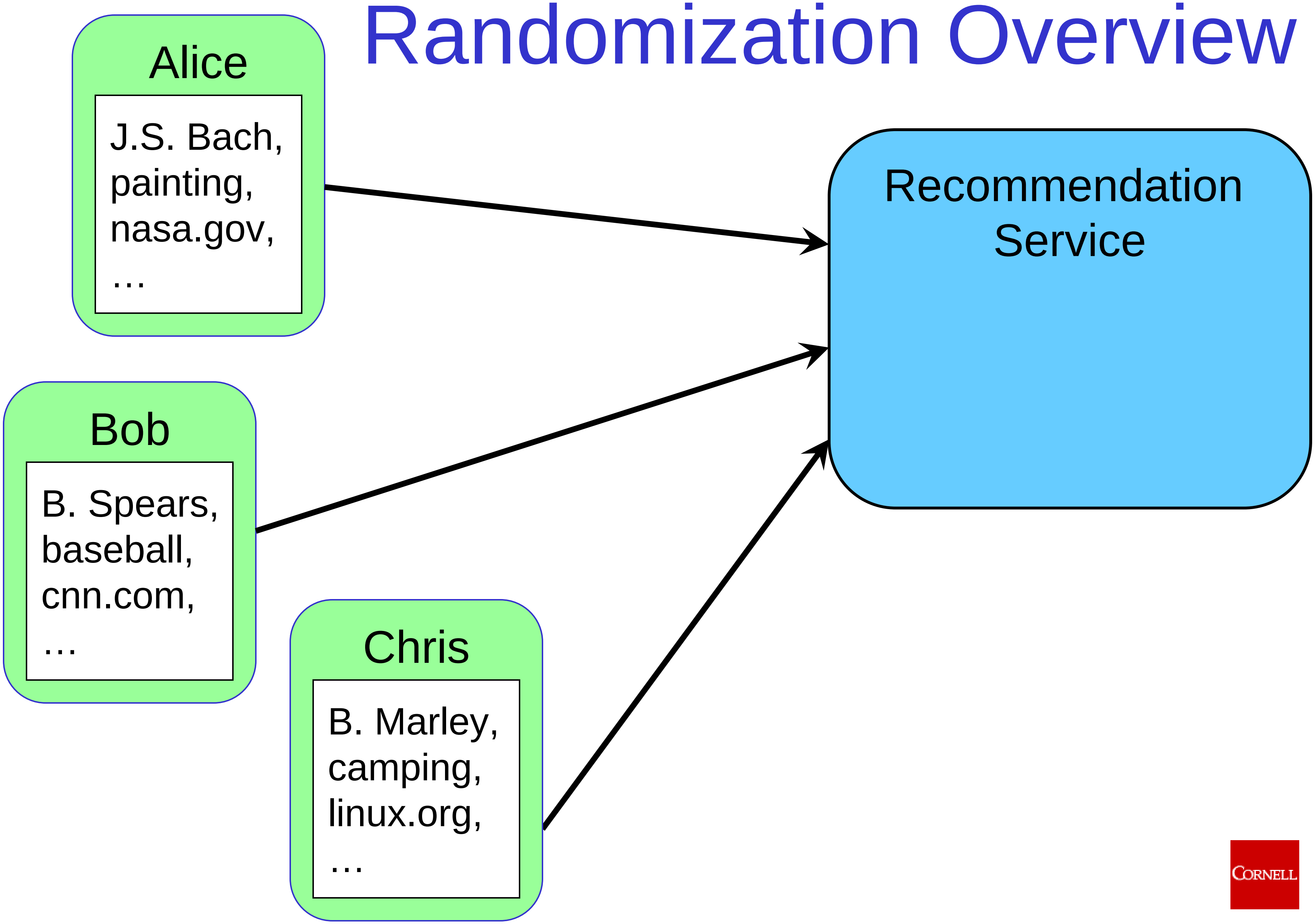
- Let $\boldsymbol{T}$ be the set of all transactions, and $\boldsymbol{t} \in \boldsymbol{T}$ be a transaction
- Any itemset $\boldsymbol{A}$ has support (frequency) $\boldsymbol{s}$ in $\boldsymbol{T}$ if

$$s = \text{supp}\,(A) = \frac{\#\{t \in T \mid A \subseteq t\}}{|T|}$$

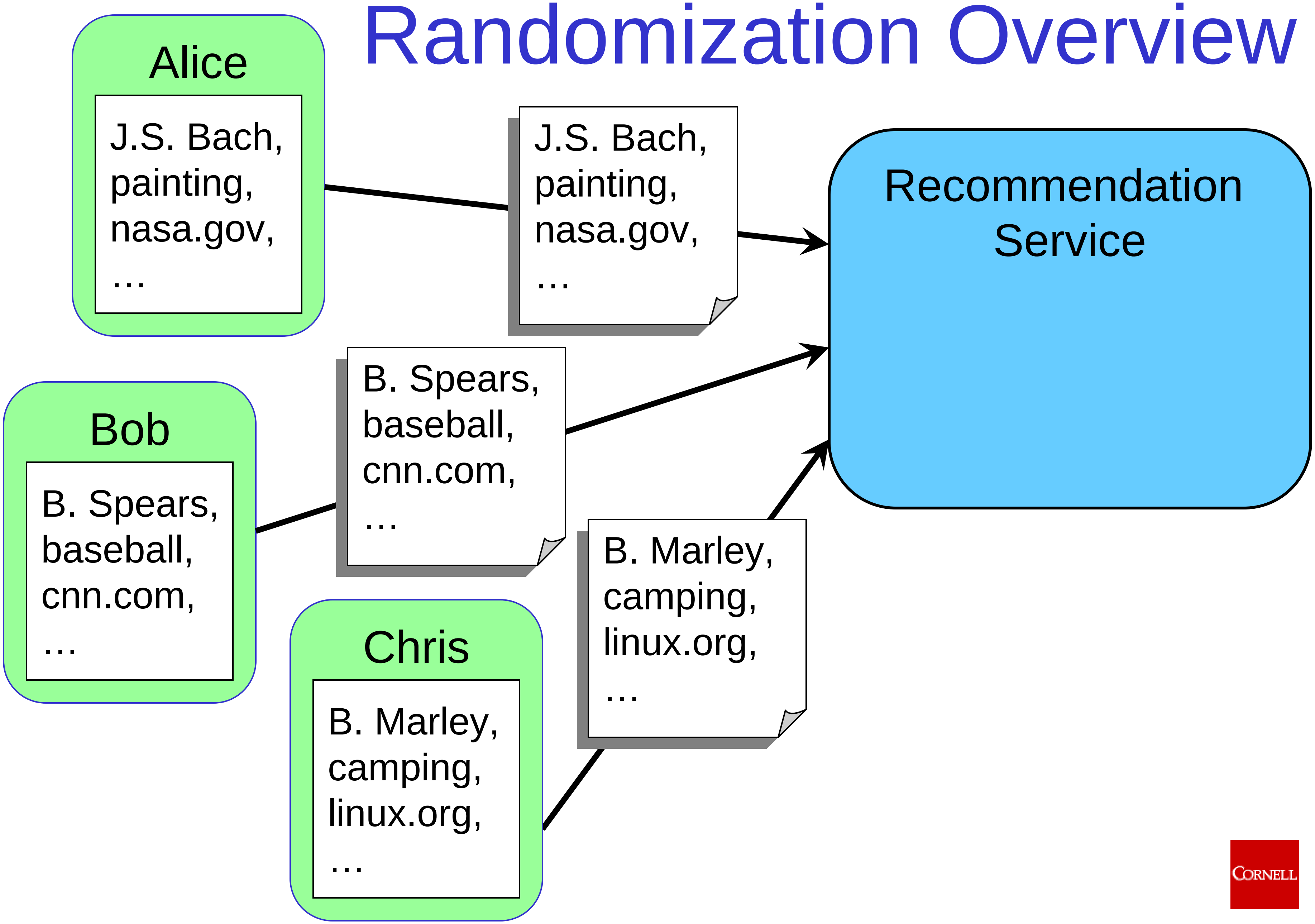
- Itemset $\boldsymbol{A}$ is frequent if $\boldsymbol{s} \geq \boldsymbol{s}_{\min}$
- Antimonotonicity: if $\boldsymbol{A} \subseteq \boldsymbol{B}$, then $\text{supp}\,(\boldsymbol{A}) \geq \text{supp}\,(\boldsymbol{B})$.
- Association rule: $\boldsymbol{A} \Rightarrow \boldsymbol{B}$ holds when the union $\boldsymbol{A} \cup \boldsymbol{B}$ is frequent and: $\text{supp}\,(\boldsymbol{A} \cup \boldsymbol{B}) \geq \text{supp}\,(\boldsymbol{A}) \cdot \boldsymbol{conf}_{\min}$

CORNELL

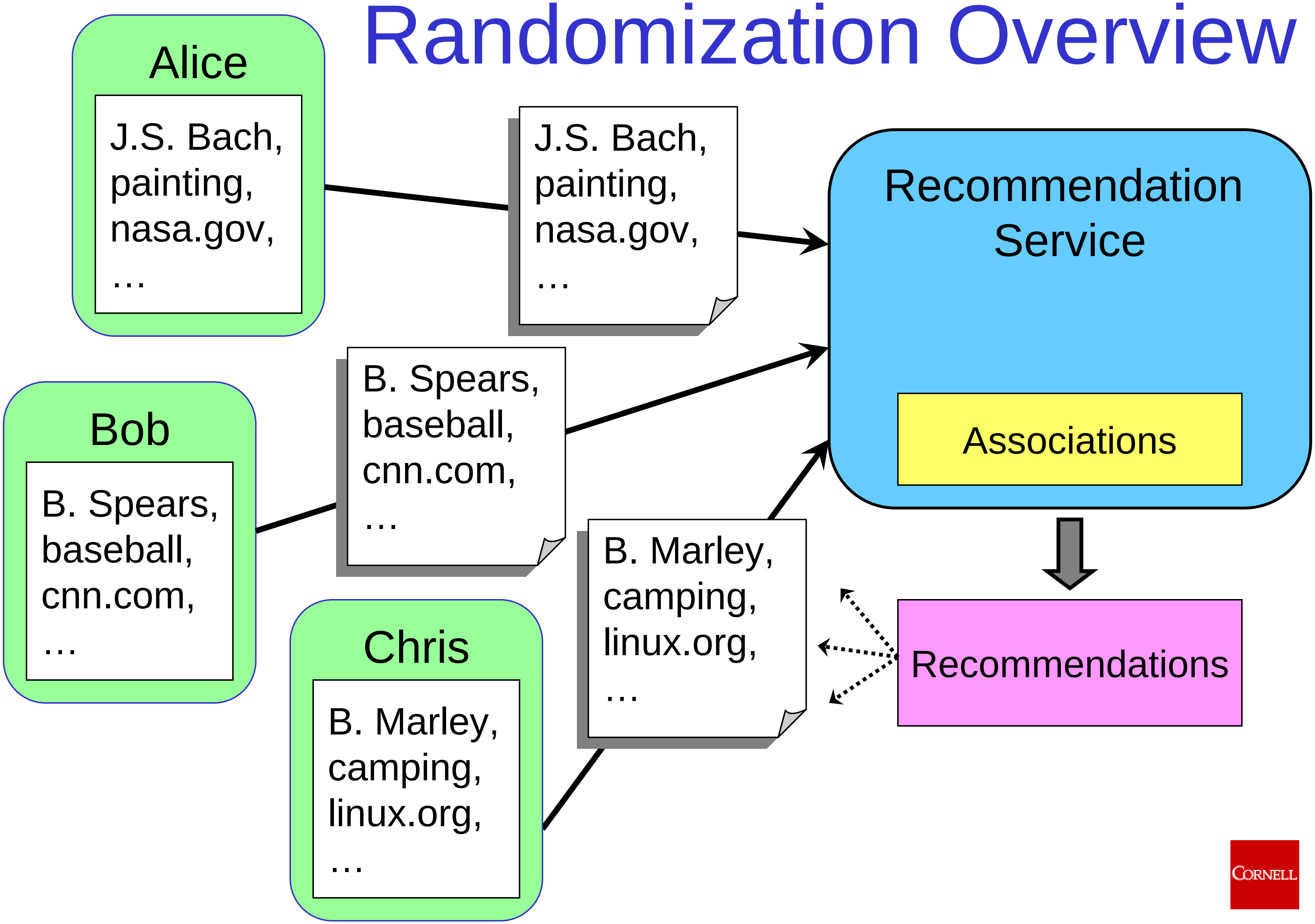
Randomization Overview



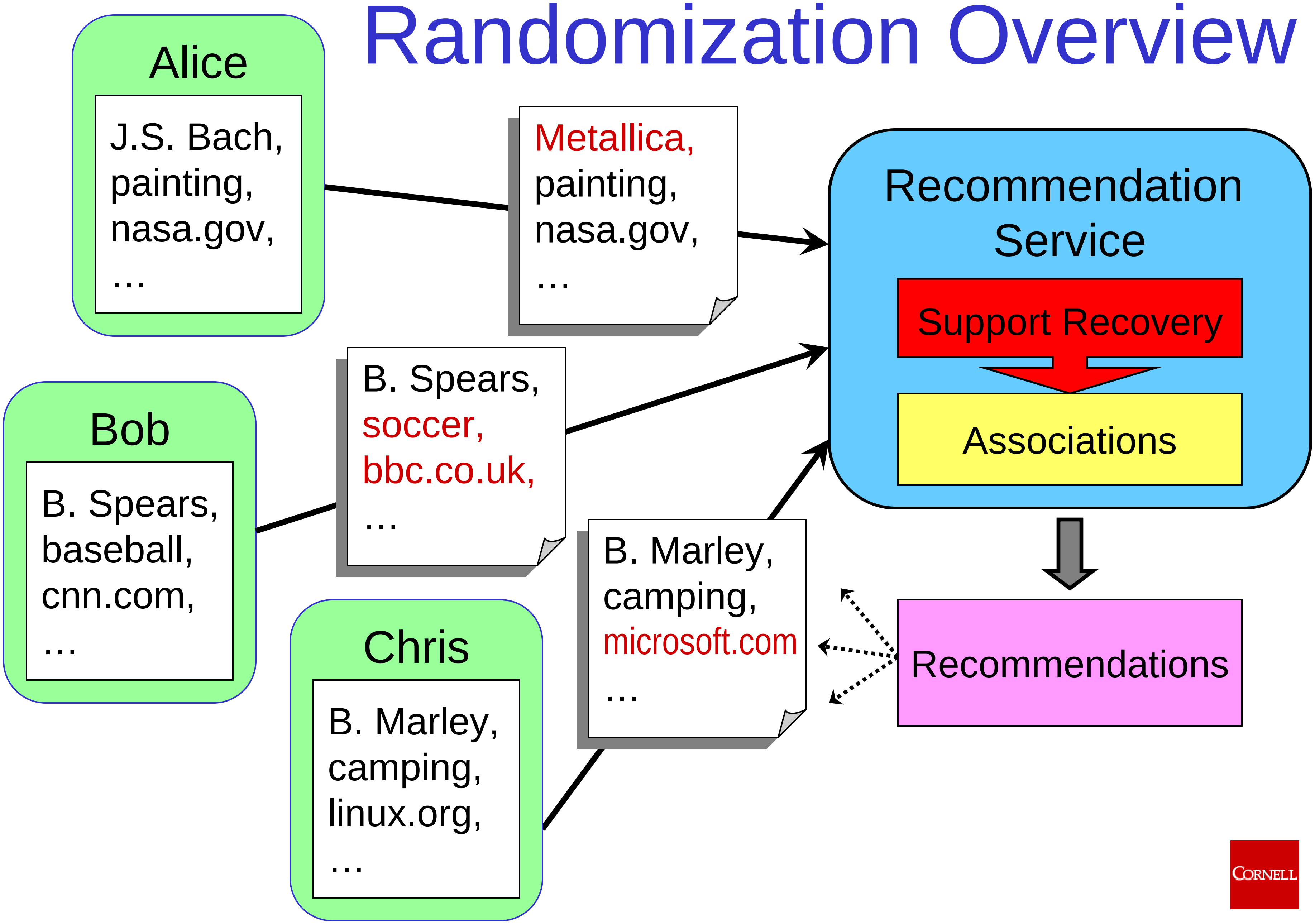
Randomization Overview



Randomization Overview



Randomization Overview



The Problem

- How to randomize transactions so that
  - we can find frequent itemsets
  - while preserving privacy at transaction level?

Randomization Example

A randomization may "look strong" but sometimes fail to hide some items of an individual transaction.

- Randomization example: given a transaction,
  - keep item with 20% probability,
  - replace with a new random item with 80% probability.

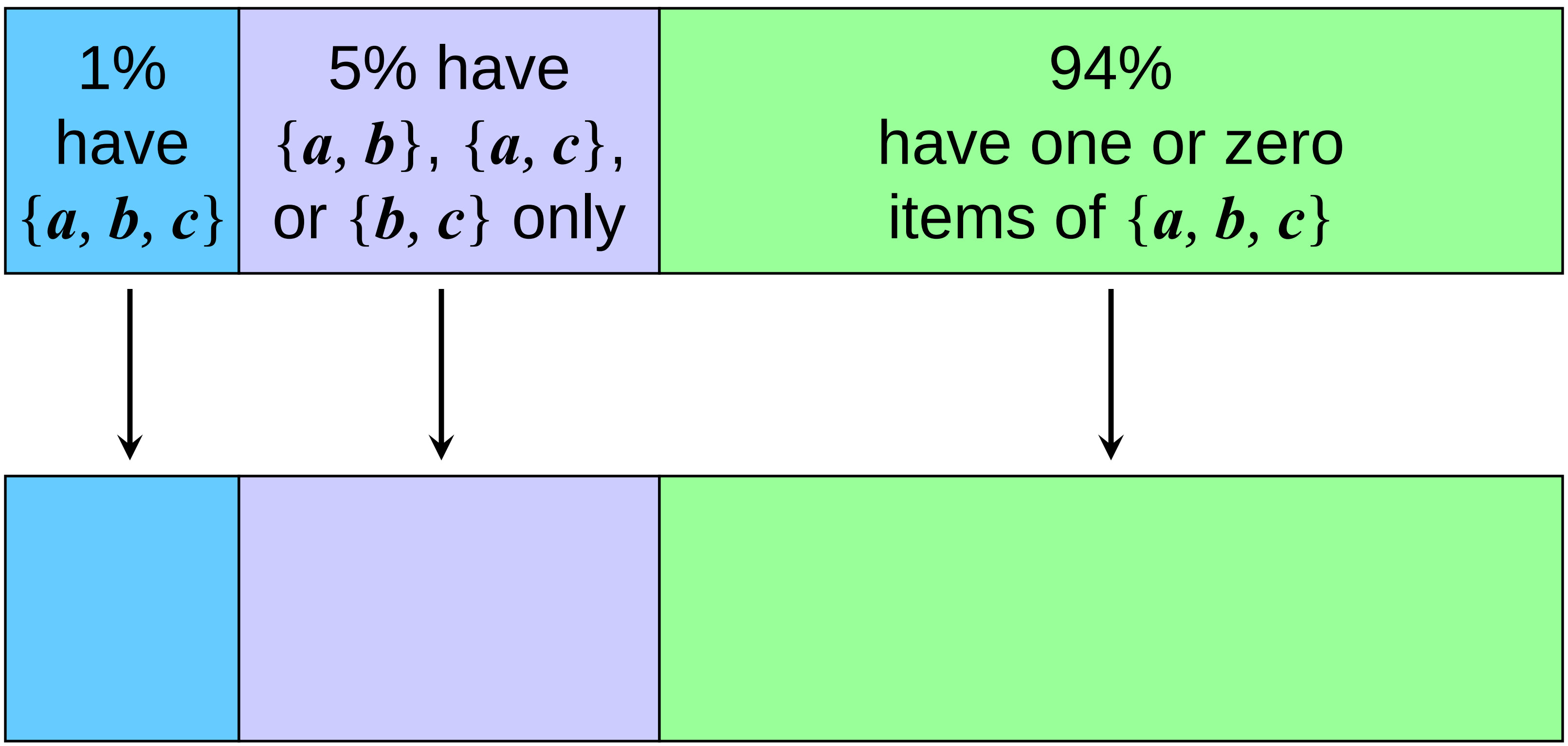
Example: $\{a, b, c\}$

10 M transactions of size 10 with 10 K items:

1% have $\{a, b, c\}$	5% have $\{a, b\}$, $\{a, c\}$, or $\{b, c\}$ only	94% have one or zero items of $\{a, b, c\}$

Example: $\{a, b, c\}$

10 M transactions of size 10 with 10 K items:



After randomization: How many have $\{a, b, c\}$?

Example: $\{a, b, c\}$

10 M transactions of size 10 with 10 K items:

1% have $\{a, b, c\}$	5% have $\{a, b\}$, $\{a, c\}$, or $\{b, c\}$ only	94% have one or zero items of $\{a, b, c\}$
$\downarrow \bullet\, 0.2^3$	$\downarrow \bullet\, 0.2^2 \bullet 8 \bullet 0.8/10{,}000$	$\downarrow$ at most $\bullet\, 0.2 \bullet (9 \bullet 0.8/10{,}000)^2$
0.008% 800 ts.	0.000128% 13 trans.	less than 0.00002% 2 transactions

After randomization: How many have $\{a, b, c\}$?

Example: $\{a, b, c\}$

10 M transactions of size 10 with 10 K items:

1% have $\{a, b, c\}$	5% have $\{a, b\}$, $\{a, c\}$, or $\{b, c\}$ only	94% have one or zero items of $\{a, b, c\}$
$\bullet\ 0.2^3$	$\bullet\ 0.2^2 \bullet 8 \bullet 0.8/10{,}000$	at most $\bullet\ 0.2 \bullet (9 \bullet 0.8/10{,}000)^2$
0.008% 800 ts. **98.2%**	0.000128% 13 trans. **1.6%**	less than 0.00002% 2 transactions **0.2%**

After randomization: How many have $\{a, b, c\}$?

Example: $\{a, b, c\}$

- Given nothing, we have only 1% probability that $\{a, b, c\}$ occurs in the original transaction
- Given $\{a, b, c\}$ in the randomized transaction, we have about 98% certainty of $\{a, b, c\}$ in the original one.
- This is what we call a privacy breach.
- The example randomization preserves privacy "on average," but not "in the worst case."

CORNELL

Privacy Breaches

- Suppose the "adversary" wants to know if $z \in t$, where
  - t is an original transaction;
  - t' is the corresponding randomized transaction;
  - A is a (frequent) itemset, $z \in A$
- Itemset A causes a privacy breach of level β (e.g. 50%) if:

$$\boxed{\text{Prob}\left[z \in t \mid A \subseteq t'\right] \geq \beta}$$

Knowledge of $A \subseteq t'$ makes a jump from $\text{Prob}\,[z \in t]$ to $\text{Prob}\,[z \in t \mid A \subseteq t']$ (in the adversary's viewpoint).

CORNELL

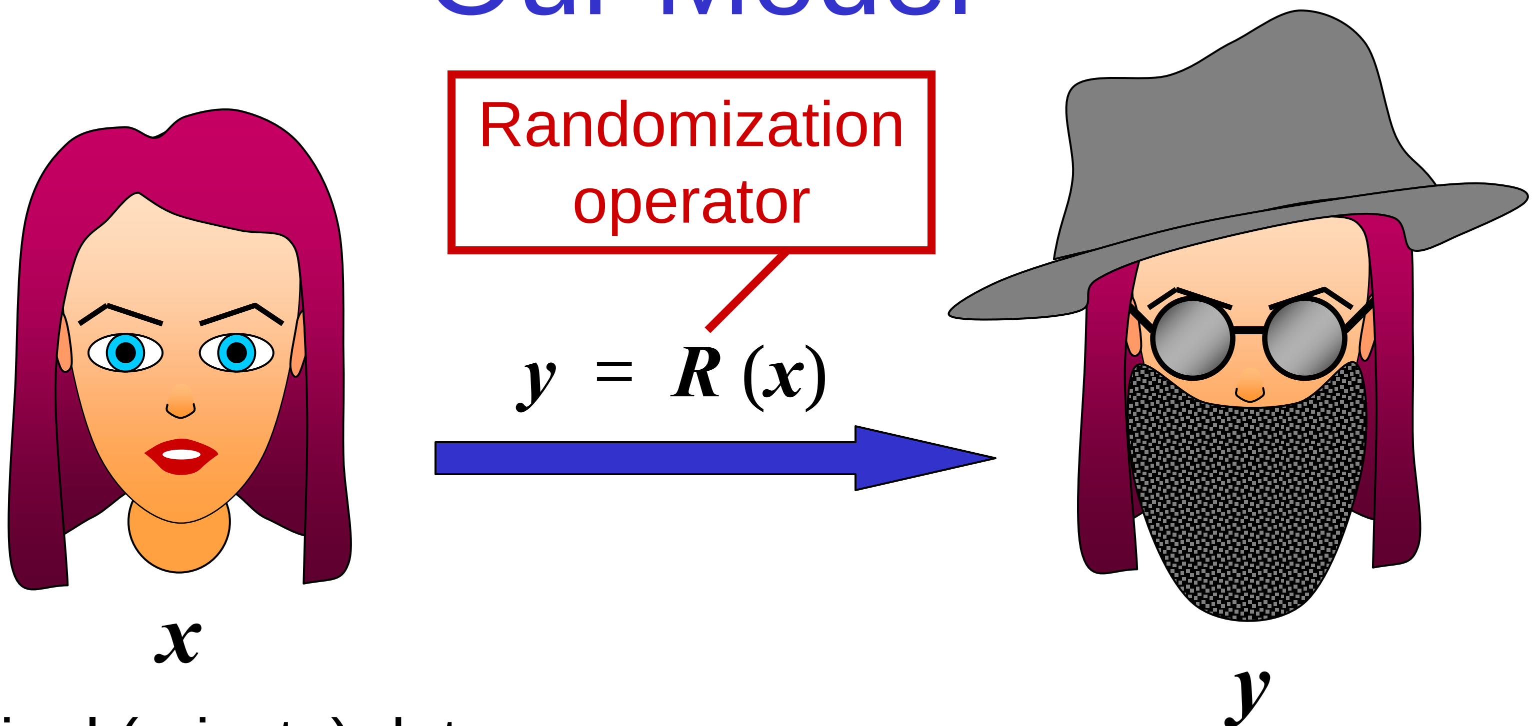
Talk Outline

- Introduction
- Randomization and privacy for association rules
- Amplification: upper bound on breaches
- Experimental results
- Conclusion

Generalized Approach

- We want a bound for all privacy breaches
  - not only for: item $\in t$ versus itemset $\subseteq t'$
- No knowledge of data distribution is required in advance
  - We don't have to know $\mathrm{Prob}\,[\text{item} \in t]$
- Applicable to numerical data as well
- Easy to work with, even for complex randomizations

Our Model



Original (private) data

Assumptions:

- Described by a random variable X.
- Each client is independent.

Randomized data

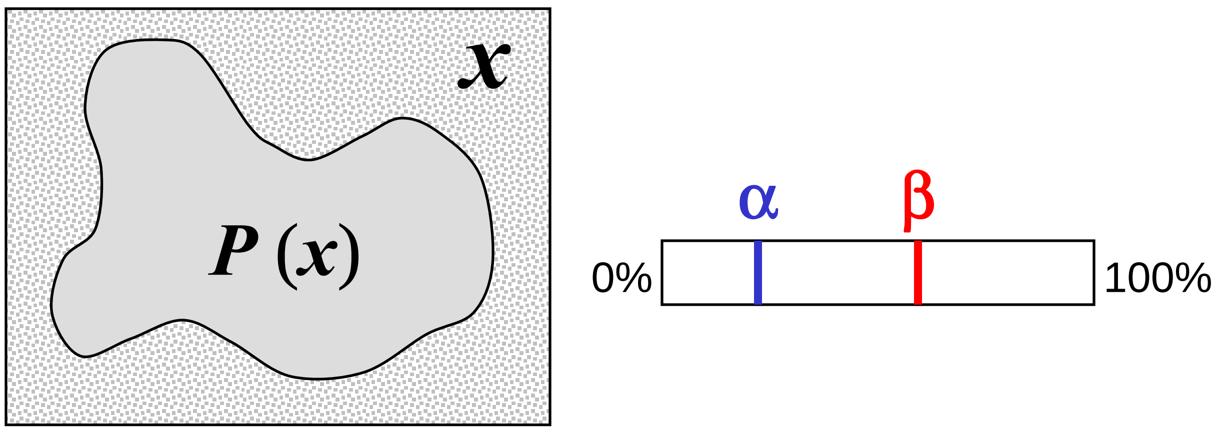
Described by a random variable $Y = R(X)$.

CORNELL

α-to-β Privacy Breach

Let $P(x)$ be any property of client's private data;

Let $0 < \alpha < \beta < 1$ be two probability thresholds.



Example:

$P(x)$ = "transaction x contains $\{a, b, c\}$"

α = 1% and β = 50%

α-to-β Privacy Breach

Let $\boldsymbol{P}(\boldsymbol{x})$ be any property of client's private data;

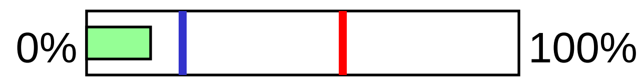
Let $0 < \alpha < \beta < 1$ be two probability thresholds.

Client

$$\boldsymbol{X} = \boldsymbol{x}$$

SERVER

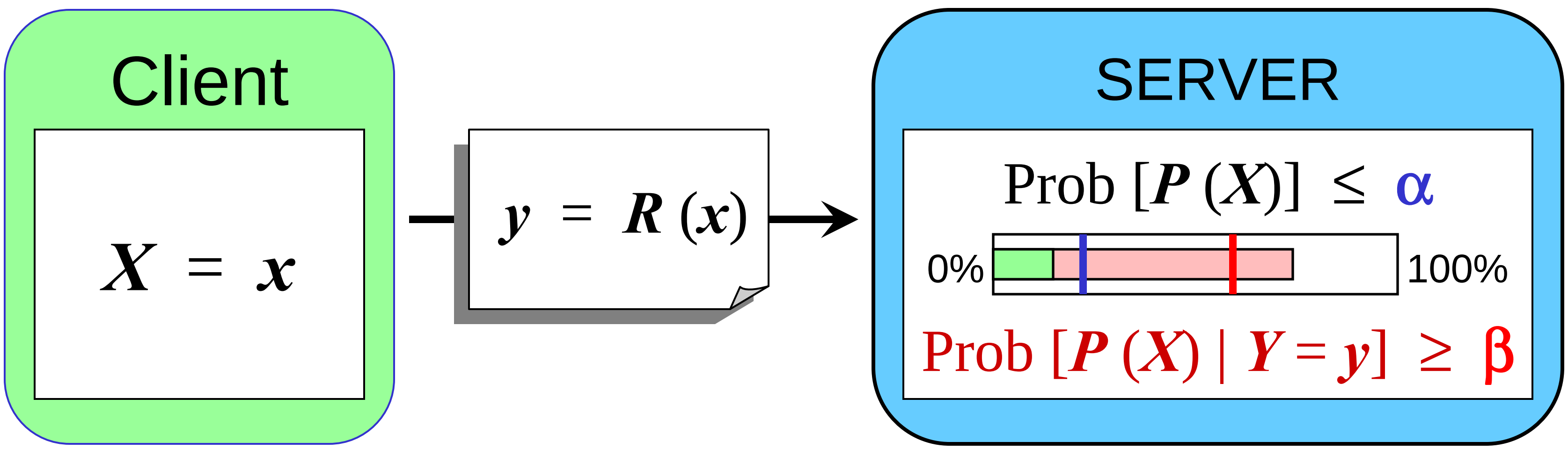
$$\text{Prob}\,[\boldsymbol{P}(\boldsymbol{X})] \le \alpha$$



α-to-β Privacy Breach

Let $\boldsymbol{P}(\boldsymbol{x})$ be any property of client's private data;

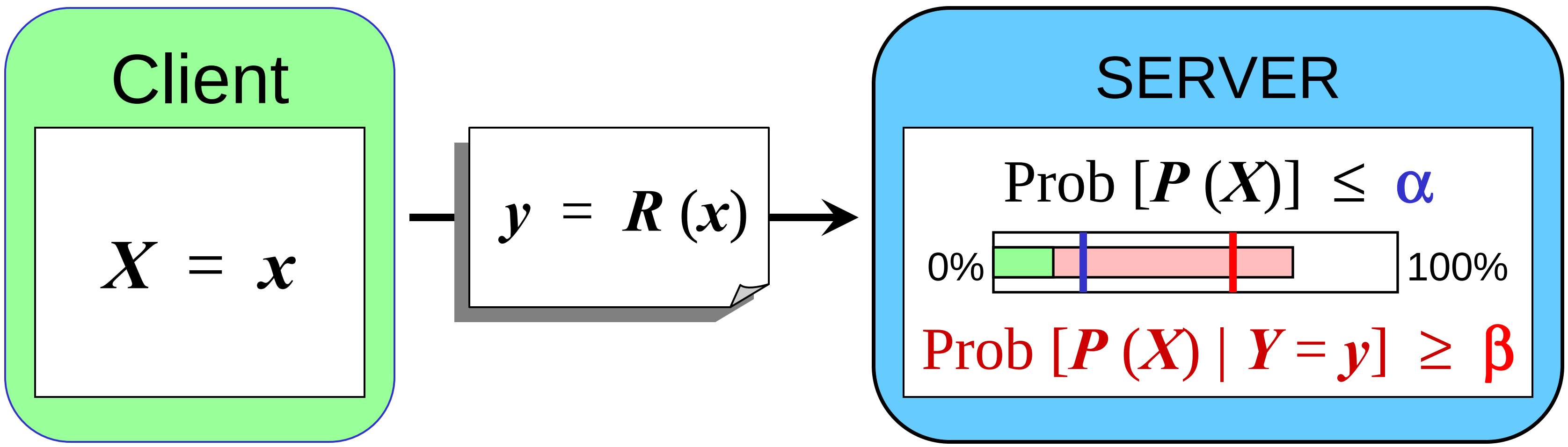
Let $0 < \alpha < \beta < 1$ be two probability thresholds.



α-to-β Privacy Breach

Let $P(x)$ be any property of client's private data;

Let $0 < \alpha < \beta < 1$ be two probability thresholds.



Disclosure of y causes an α-to-β privacy breach w.r.t. property $P(x)$.

α-to-β Privacy Breach

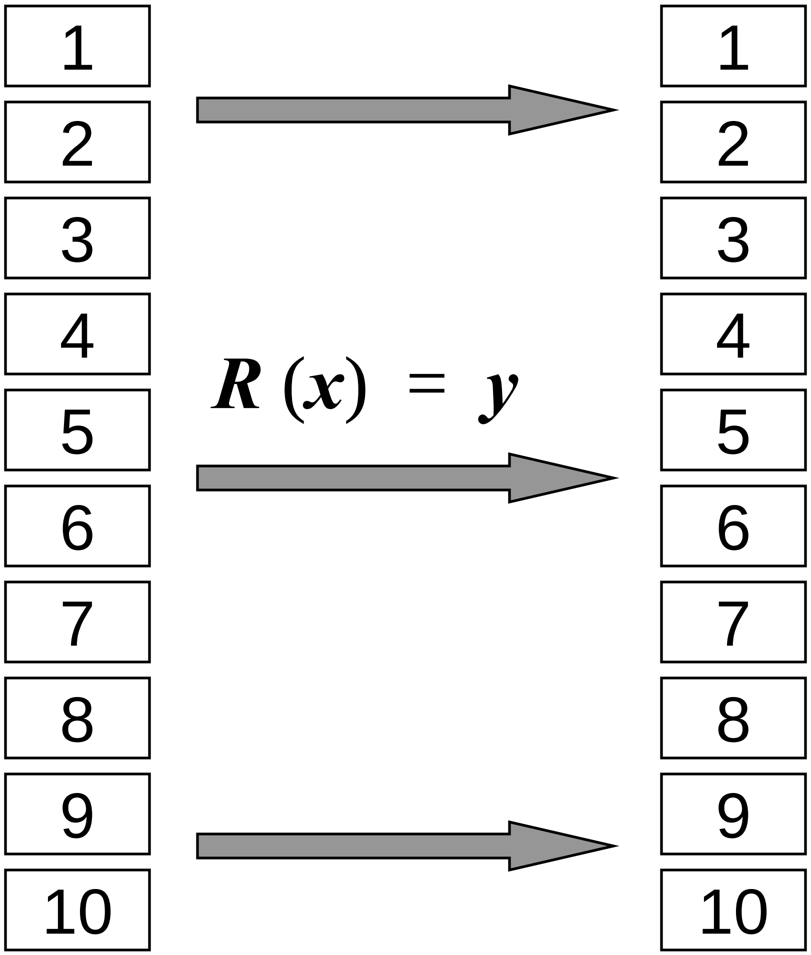
Checking for α-to-β privacy breaches:

- There are exponentially many properties $\boldsymbol{P}(\boldsymbol{x})$;
- We have to know the data distribution in order to check whether $\text{Prob}\,[\boldsymbol{P}(\boldsymbol{X})] \leq \alpha$ and $\text{Prob}\,[\boldsymbol{P}(\boldsymbol{X}) \mid \boldsymbol{Y} = \boldsymbol{y}] \geq \beta$.

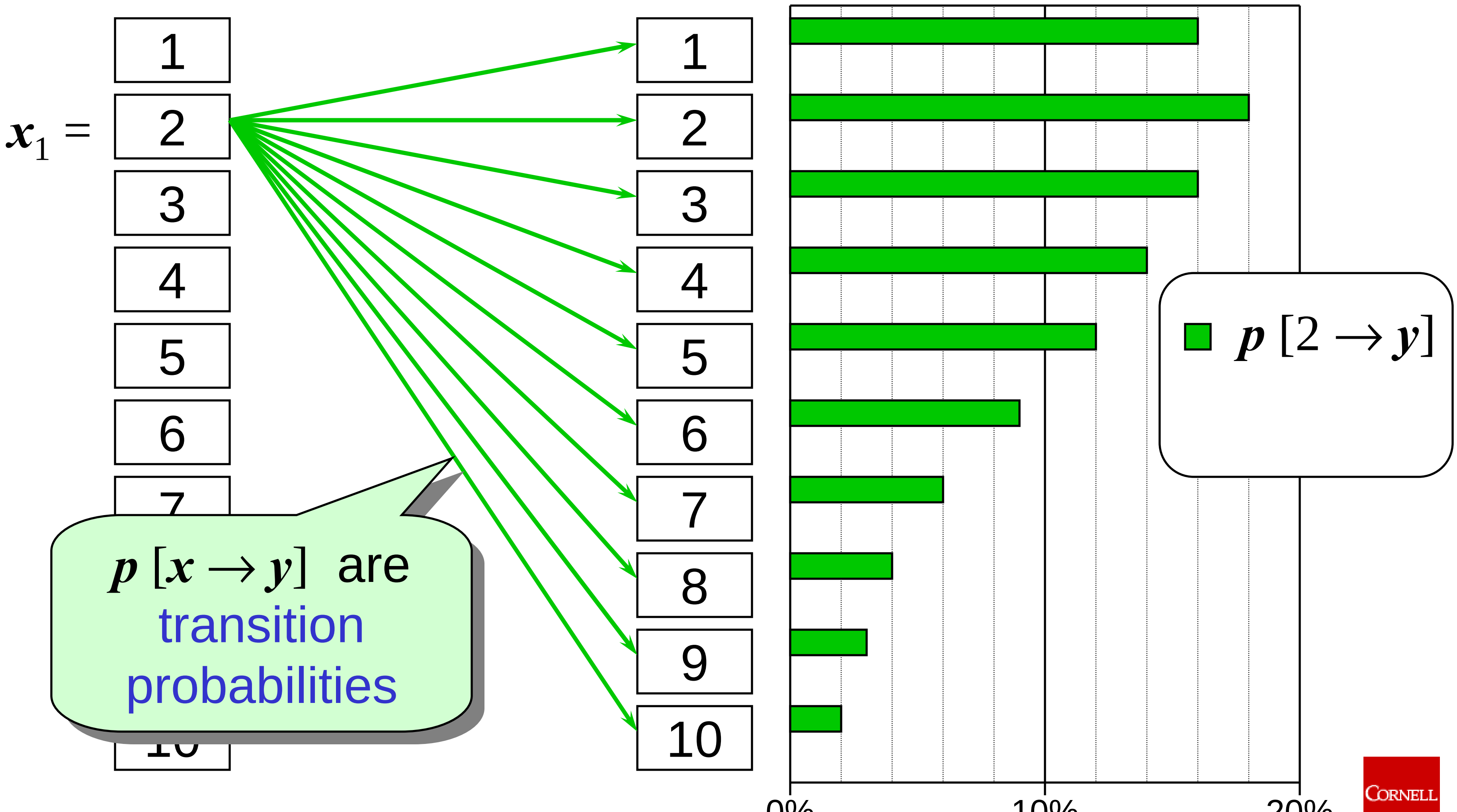
Is there a simple property of randomization operator $\boldsymbol{R}$ that limits privacy breaches?

CORNELL

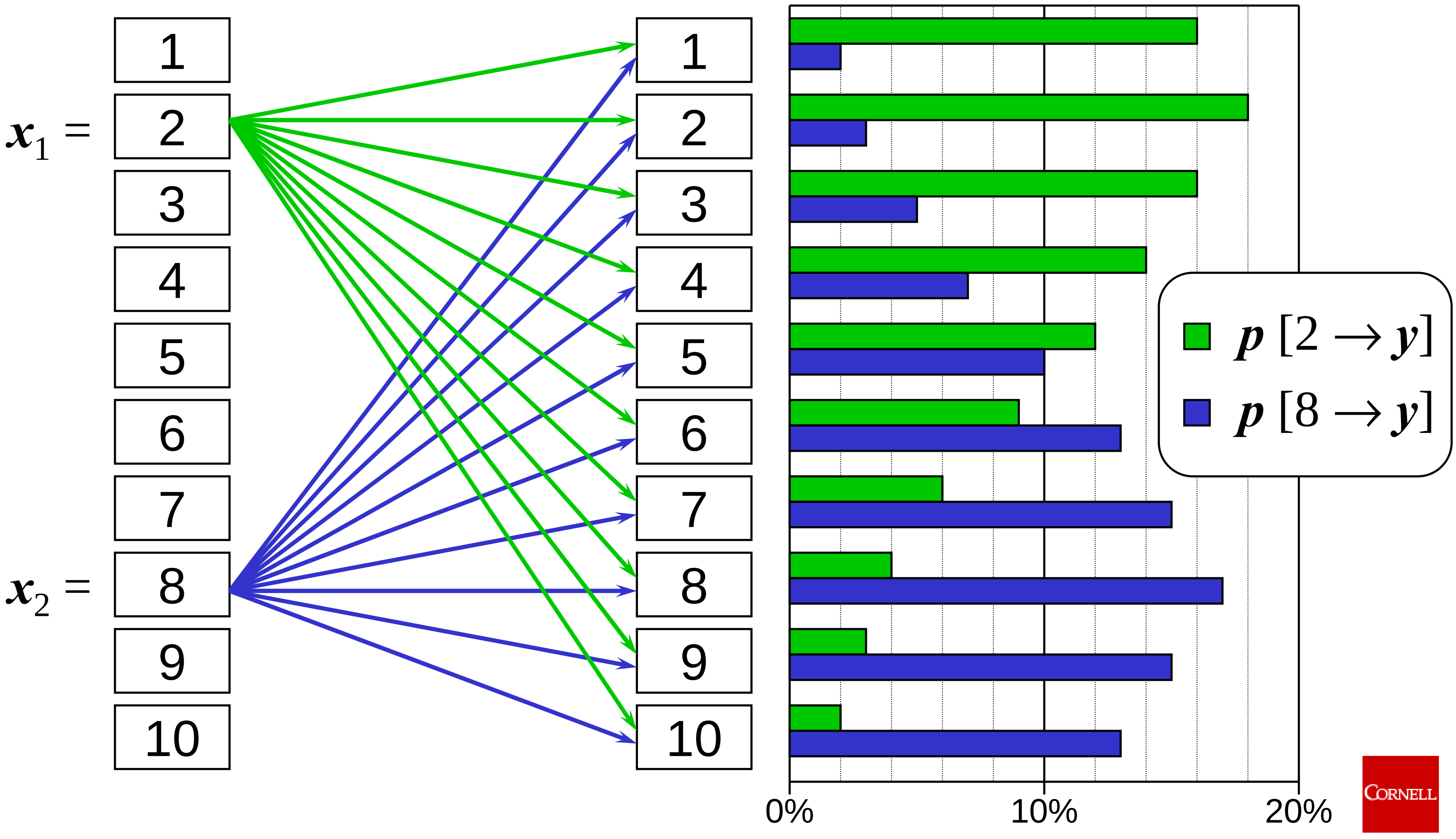
Amplification Condition



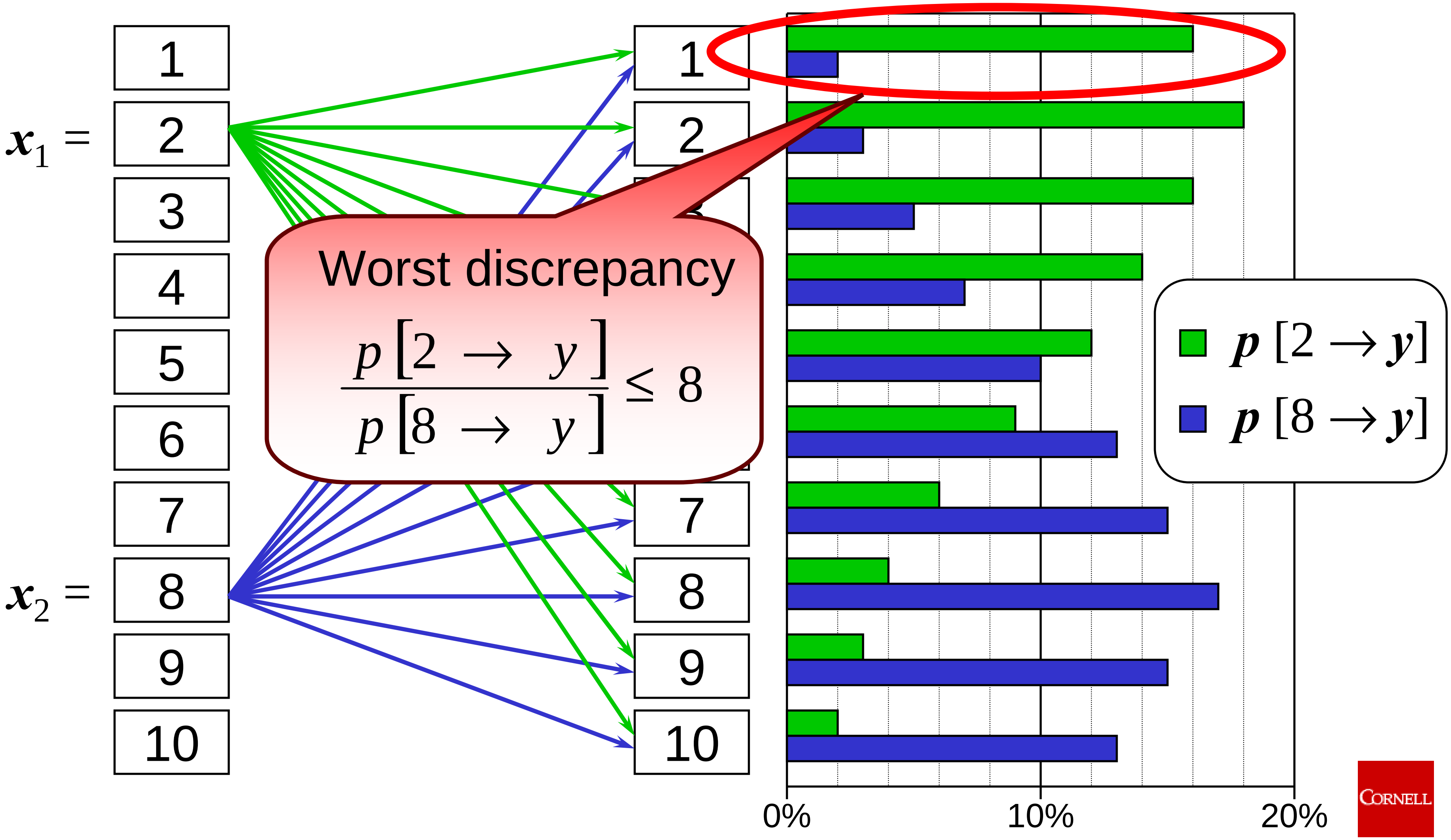
Amplification Condition



Amplification Condition



Amplification Condition



CORNELL

Amplification Condition

Definition:

- Randomization operator $\boldsymbol{R}$ is called "at most γ-amplifying" if:

$$\max_{x_1, x_2} \max_{y} \frac{p[x_1 \to y]}{p[x_2 \to y]} \leq \gamma$$

  - Transition probabilities $\boldsymbol{p}\ [\boldsymbol{x} \to \boldsymbol{y}] = \text{Prob}\ [\boldsymbol{R}\ (\boldsymbol{x}) = \boldsymbol{y}]$ depend only on the operator $\boldsymbol{R}$ and not on data.
  - We assume that all $\boldsymbol{y}$ have a nonzero probability.
- The bigger γ is, the more may be revealed about $\boldsymbol{x}$.

CORNELL

The Bound on α-to-β Breaches

Statement:

- If randomization operator $\boldsymbol{R}$ is at most γ-amplifying, and if:

$$\gamma < \frac{\beta}{\alpha} \cdot \frac{1-\alpha}{1-\beta}$$

- Then, revealing $\boldsymbol{R}(\boldsymbol{X})$ to the server will never cause an α-to-β privacy breach.

See proof in [PODS 2003].

CORNELL

The Bound on α-to-β Breaches

Examples:

- 5%-to-50% privacy breaches do not occur for $\gamma < 19$:

$$\frac{0.5}{0.05} \cdot \frac{1 - 0.05}{1 - 0.5} = 19$$

- 1%-to-98% privacy breaches do not occur for $\gamma < 4851$:

$$\frac{0.98}{0.01} \cdot \frac{1 - 0.01}{1 - 0.98} = 4851$$

- 50%-to-100% privacy breaches do not occur for any finite γ.

Amplification: Summary

- An α-to-β privacy breach w.r.t. property $\boldsymbol{P}(\boldsymbol{x})$ occurs when
  - Prob [$\boldsymbol{P}$ is true] $\leq \alpha$
  - Prob [$\boldsymbol{P}$ is true | $\boldsymbol{Y} = \boldsymbol{y}$] $\geq \beta$.
- Amplification methodology limits privacy breaches by just looking at transitional probabilities of randomization.
  - Does not use data distribution:

$$\max_{x_1, x_2} \max_{y} \frac{p[x_1 \to y]}{p[x_2 \to y]} \leq \gamma$$

Amplification In Practice

- Given transaction t of size m, construct $t' = R(t)$:

$t = \boxed{a, b, c, d, e, f, u, v, w}$

$t' =$

Definition of select-a-size

- Given transaction $\boldsymbol{t}$ of size $\boldsymbol{m}$, construct $\boldsymbol{t'} = \boldsymbol{R}\,(\boldsymbol{t})$:
  - Choose a number $\boldsymbol{j} \in \{0, 1, \ldots, \boldsymbol{m}\}$ with distribution $\{\boldsymbol{p}\,[\boldsymbol{j}]\}_{0..\boldsymbol{m}}$;

$$\boldsymbol{t} = \boxed{a, b, c, d, e, f, u, v, w}$$

$$\boldsymbol{t'} = \underbrace{\boxed{}}_{j\,=\,4}$$

CORNELL

Definition of select-a-size

- Given transaction $\boldsymbol{t}$ of size $\boldsymbol{m}$, construct $\boldsymbol{t'} = \boldsymbol{R}(\boldsymbol{t})$:
  - Choose a number $j \in \{0, 1, \ldots, m\}$ with distribution $\{p[j]\}_{0..m}$;
  - Include exactly j items of t into t';

$$t = \boxed{a, \color{red}{b}, c, d, \color{red}{e}, f, \color{red}{u}, v, \color{red}{w}}$$

$$t' = \underbrace{\boxed{\color{red}{b}, \color{red}{e}, \color{red}{u}, \color{red}{w}}}_{j\,=\,4}$$

Definition of select-a-size

- Given transaction $\boldsymbol{t}$ of size $\boldsymbol{m}$, construct $\boldsymbol{t'} = \boldsymbol{R}(\boldsymbol{t})$:
  - Choose a number $\boldsymbol{j} \in \{0, 1, \ldots, \boldsymbol{m}\}$ with distribution $\{\boldsymbol{p}[\boldsymbol{j}]\}_{0..\boldsymbol{m}}$;
  - Include exactly $\boldsymbol{j}$ items of $\boldsymbol{t}$ into $\boldsymbol{t'}$;
  - Each other item (not from $\boldsymbol{t}$) goes into $\boldsymbol{t'}$ with probability ρ.

The choice of $\{\boldsymbol{p}[\boldsymbol{j}]\}_{0..\boldsymbol{m}}$ and ρ is based on the desired privacy level.

$\boldsymbol{t}$ = ***a, b, c, d, e, f, u, v, w***

$\boldsymbol{t'}$ = ***b, e, u, w*** | ***œ, å, ß, §, ψ, €, ℵ, ъ, ђ, …***

$\boldsymbol{j} = 4$ ⟷ items inserted with prob. ρ

Talk Outline

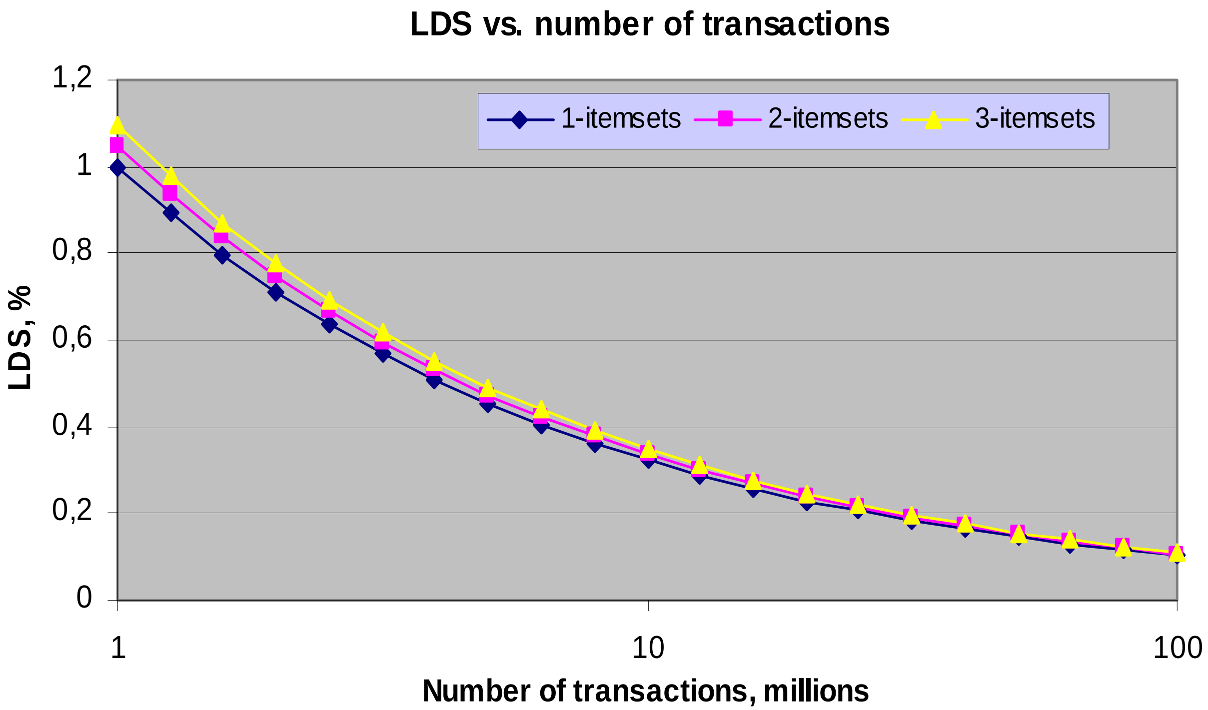
- Introduction
- Randomization and privacy for association rules
- Amplification: upper bound on breaches
- Experimental results
- Conclusion

CORNELL

Lowest Discoverable Support

- LDS is s.t., when predicted, is 4σ away from zero.
- Roughly, LDS is proportional to $1 \big/ \sqrt{\#\ \text{trans.}}$

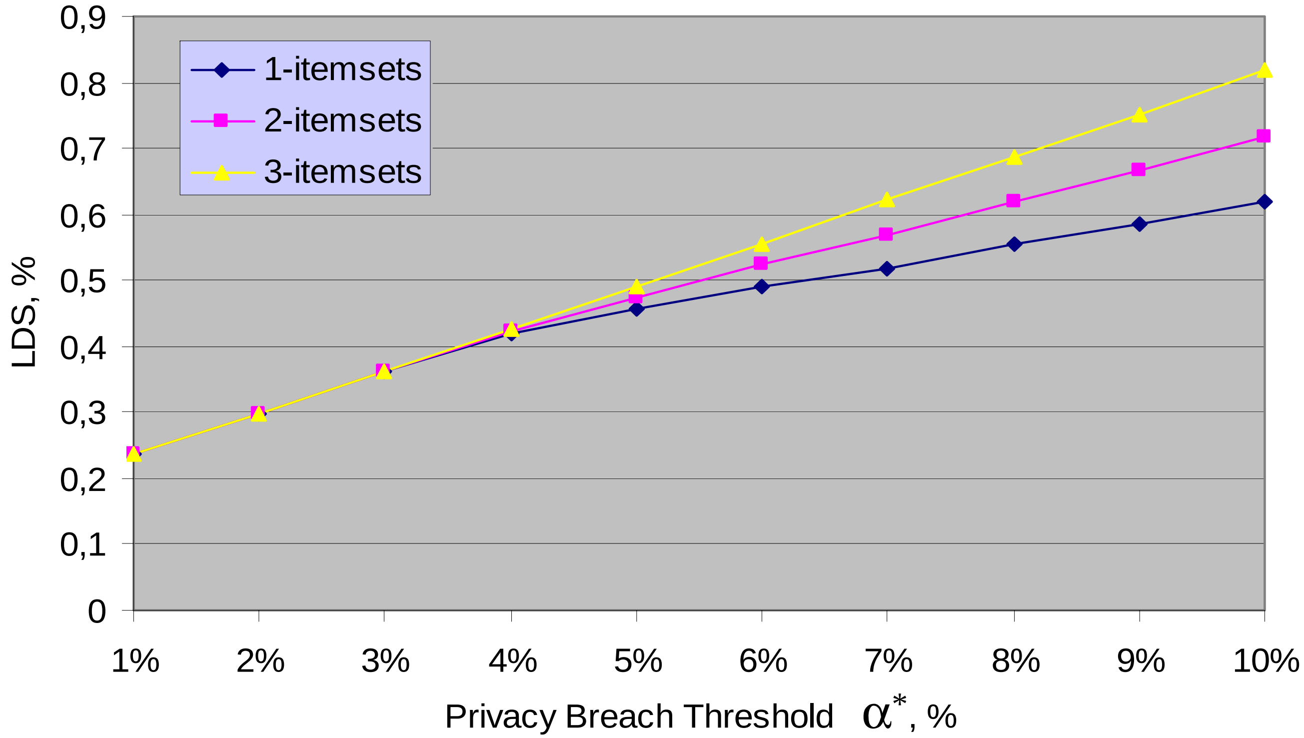
$|\, t \,| = 5$, 5%-50% privacy breaches are the worst allowed



CORNELL

LDS vs. Breach Threshold α^*

$|\,\boldsymbol{t}\,| = 5$, 5 M trans., α-to-50% privacy breaches are limited to $\alpha \geq \alpha^*$

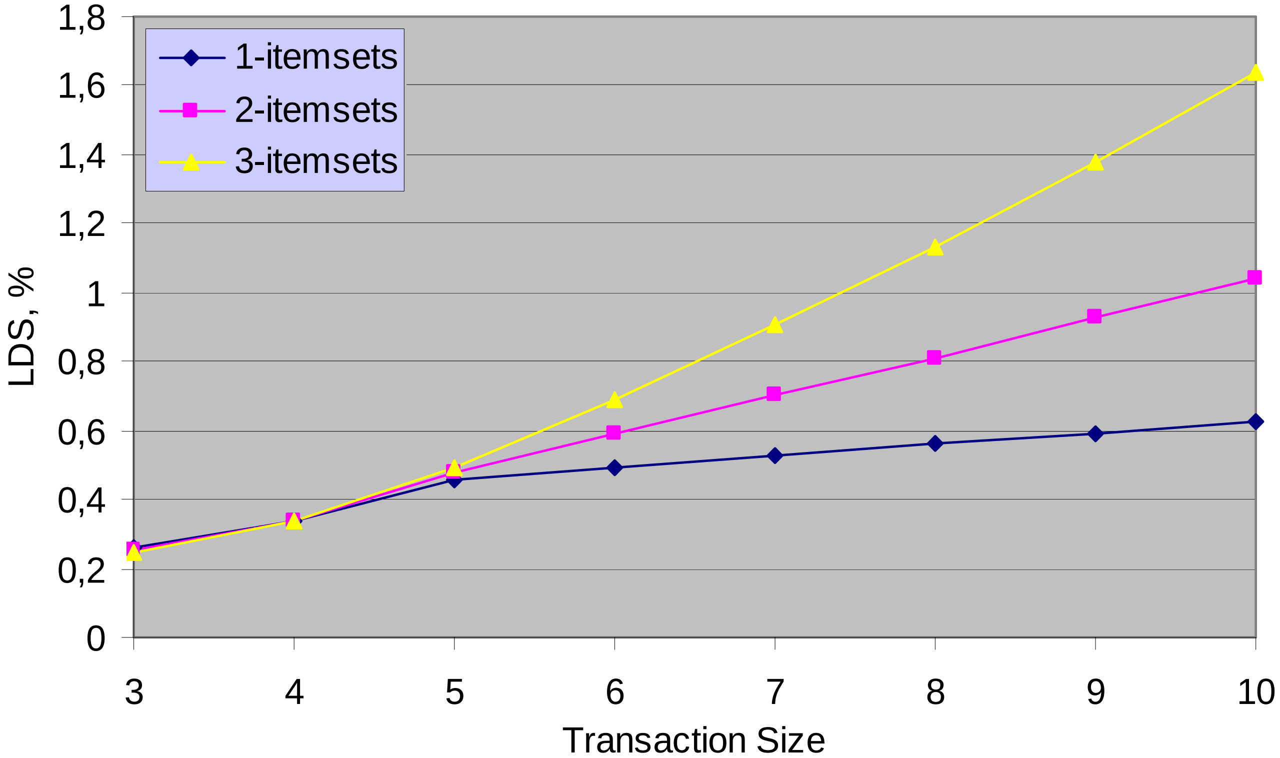


- Reminder: α-to-50% breach occurs when $\text{Prob}\,[\boldsymbol{P}\,(\boldsymbol{t})] \leq \alpha$ and $\text{Prob}\,[\boldsymbol{P}\,(\boldsymbol{t})\,|\,\boldsymbol{R}\,(\boldsymbol{t}) = \boldsymbol{t'}] \geq$ 50%.

CORNELL

LDS vs. Transaction Size

5%-50% privacy breaches are the worst allowed, $|\boldsymbol{T}| = 5$ M



- Longer transactions are harder to use in support recovery

CORNELL

Real datasets: soccer, mailorder

- Soccer is the clickstream log of WorldCup’98 web site, split into sessions of HTML requests.
  - 11 K items (HTMLs), 6.5 M transactions
  - Available at http://www.acm.org/sigcomm/ITA/
- Mailorder is a purchase dataset from a certain on-line store
  - Products are replaced with their categories
  - 96 items (categories), 2.9 M transactions

A small fraction of transactions are discarded as too long.

- longer than 10 (for soccer) or 7 (for mailorder)

Restricted Privacy Breaches

- Real data experiments used older approach [KDD 2002]
  - We constrained only $z \in t$ versus $A \subseteq t'$ privacy breaches
  - Restrictions in the form $\mathrm{Prob}\,[z \in t \mid A \subseteq t'] < \beta$
  - Older approach used some (minimal) information about data distribution to choose randomization parameters

Modified Apriori on Real Data

Breach level $\beta = 50\%$. Inserted 20-50% items to each transaction.

Soccer:

$s_{min} = 0.2\%$

$\sigma \approx 0.07\%$ for 3-itemsets

Itemset Size	True Itemsets	True Positives	False Drops	False Positives
1	266	254	12	31
2	217	195	22	45
3	48	43	5	26

Mailorder:

$s_{min} = 0.2\%$

$\sigma \approx 0.05\%$ for 3-itemsets

Itemset Size	True Itemsets	True Positives	False Drops	False Positives
1	65	65	0	0
2	228	212	16	28
3	22	18	4	5

CORNELL

False Drops   False Positives

Soccer

Pred. supp%, when true supp ≥ **0.2%**

Size	< 0.1	0.1-0.15	0.15-0.2	≥0.2
1	0	2	10	254
2	0	5	17	195
3	0	1	4	43

True supp%, when pred. supp ≥ **0.2%**

Size	< 0.1	0.1-0.15	0.15-0.2	≥0.2
1	0	7	24	254
2	7	10	28	195
3	5	13	8	43

Mailorder

Pred. supp%, when true supp ≥ **0.2%**

Size	< 0.1	0.1-0.15	0.15-0.2	≥0.2
1	0	0	0	65
2	0	1	15	212
3	0	1	3	18

True supp%, when pred. supp ≥ **0.2%**

Size	< 0.1	0.1-0.15	0.15-0.2	≥0.2
1	0	0	0	65
2	0	0	28	212
3	1	2	2	18

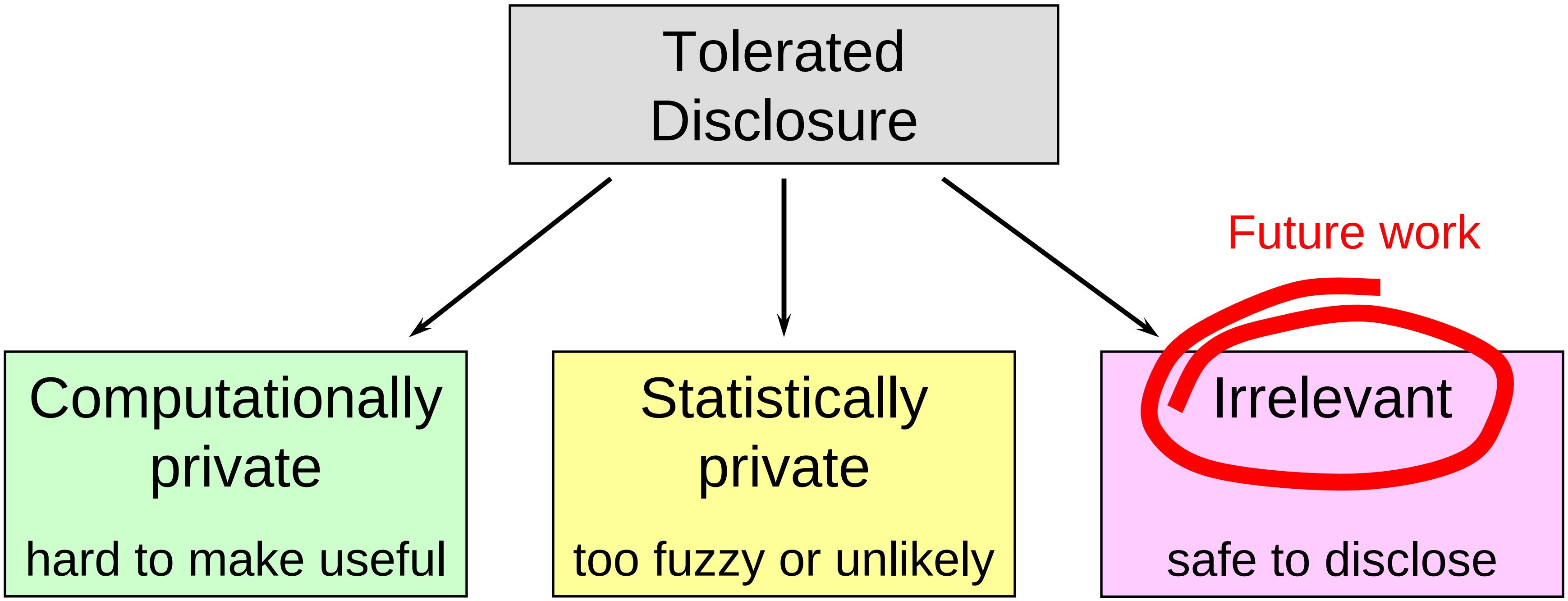
Actual Privacy Breaches

- Verified actual privacy breach levels
- The breach probabilities $\mathrm{Prob}\,[z \in t \mid A \subseteq t']$ are counted in the datasets for frequent and near-frequent itemsets.
- With the right choice of randomization parameters, even worst-case breach levels fluctuated around 50%
  - At most 53.2% for soccer,
  - At most 55.4% for mailorder.

Ongoing Research

- Using randomization and traditional secure multiparty computation together
  - Privacy preserving two-party join size computation with sketches
- What if we cannot guarantee amplification condition?
  - Probabilistic privacy breaches and amplification "on average"
- Information theory and statistical privacy
  - A slightly modified information measure that provably bounds privacy breaches

Future Work



- Can statistical privacy be extended so that we can prove "orthogonality" between disclosure and sensitive questions?

Conclusion

- We defined privacy using statistics, not computational hardness
- Randomization can guarantee statistical privacy
  - Demonstrated for association mining
- A simple property of randomization operators provably bounds privacy breaches

CORNELL

Thank You!

Questions?